(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 1 074 617 A2

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
07.02.2001 Bulletin 2001/06

(21) Application number: **00116126.4**

(22) Date of filing: **28.07.2000**

(51) Int Cl.[7]: **C12N 15/12**, C12N 15/11, C07K 14/47, C07K 16/18, C12Q 1/68

(84) Designated Contracting States:
AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE
Designated Extension States:
AL LT LV MK RO SI

(30) Priority: **29.07.1999 JP 24803699**
27.08.1999 JP 30025399
11.01.2000 JP 2000118776
02.05.2000 JP 2000183767
09.06.2000 JP 2000241899

(71) Applicant: **Helix Research Institute**
Kisarazu-shi, Chiba 292-0812 JP

(72) Inventors:
- **Ota, Toshio**
  Fujisawa-shi, Kanagawa 251-0042 JP
- **Isogai, Takao**
  Inashiki-gun, Ibaraki 300-0303 JP
- **Nishikawa, Tetsuo**
  Itabashi-ku, Tokyo 173-0013 JP
- **Hayashi, Kohji**
  Ichihara-shi, Chiba 292-0056 JP
- **Saito, Kaoru**
  Kisarazu-shi, Chiba 292-0056 JP
- **Yamamoto, Junichi**
  Kisarazu-shi, Chiba 292-0041 JP
- **Ishii, Shizuko**
  Kisarazu-shi, Chiba 292-0812 JP
- **Sugiyama, Tomoyasu**
  Kisarazu-shi, Chiba 292-0045 JP
- **Wakamatsu, Ai**
  Kisarazu-shi, Chiba 292-0014 JP
- **Nagai, Keiichi**
  Higashiyamato-shi, Tokyo 207-0022 JP
- **Otsuki, Tetsuji**
  Kisarazu-shi, Chiba 292-0045 JP

(74) Representative: **VOSSIUS & PARTNER**
Siebertstrasse 4
81675 München DE

(54) Primers for synthesising full-length cDNA and their use

(57) Primers for synthesizing full-length cDNAs and their use are provided.

5602 cDNA encoding a human protein has been isolated and nucleotide sequences of 5'-, and 3' -ends of the cDNA have been determined. Furthermore, primers for synthesizing the full-length cDNA have been provided to clarify the function of the protein encoded by the cDNA. The full-length cDNA of the present invention containing the translation start site provides information useful for analyzing the functions of the protein.

EP 1 074 617 A2

Description

FIELD OF THE INVENTION

[0001] The present invention relates to a polynucleotide encoding a novel protein, a protein encoded by the polynucleotide, and new uses of these.

BACKGROUND OF THE INVENTION

[0002] Currently, the sequencing projects, the determination and analysis of the genomic DNA of various living organisms have been in progress all over the world. The whole genomic sequences of more than 10 species of prokaryotes, a lower eukaryote, yeast, and a multicellular eukaryote, C. elegans are already determined. As to human genome, which is supposed to be composed of three thousand million base pairs, the world wide cooperative projects have been under way to analyze it, and the whole structure is predicted to be determined by the years 2002-2003. The aim of the determination of genomic sequence is to reveal the functions of all genes and their regulation and to understand living organisms as a network of interactions between genes, proteins, cells or individuals through deducing the information in a genome, which is a blueprint of the highly complicated living organisms. To understand living organisms by utilizing the genomic information from various species is not only important as an academie subject, but also socially significant from the viewpoint of industrial application.

[0003] However, determination of genomic sequences itself cannot identify the functions of all genes. For example, as for yeast, only the function of approximately half of the 6000 genes, which is predicted based on the genomic sequence, was able to be deduced. As for human, the number of the genes is predicted to be approximately one hundred thousand. Therefore, it is desirable to establish "a high throughput analysis system of the gene functions" which allows us to identify rapidly and efficiently the functions of vast amounts of the genes obtained by the genomic sequencing.

[0004] Many genes in the eukaryotic genome are split by introns into multiple exons. Thus, it is difficult to predict correctly the structure of encoded protein solely based on genomic information. In contrast, cDNA, which is produced from mRNA that lacks introns, encodes a protein as a single continuous amino acid sequence and allows us to identify the primary structure of the protein easily. In human cDNA research, to date, more than one million ESTs (Expression Sequence Tags) are publicly available, and the ESTs presumably cover not less than 80% of all human genes.

[0005] The information of ESTs is utilized for analyzing the structure of human genome, or for predicting the exon-regions of genomic sequences or their expression profile. However, many human ESTs have been derived from proximal regions to the 3'-end of cDNA, and information around the 5'-end of mRNA is extremely little. Among these human cDNAs, the number of the corresponding mRNAs whose encoding protein sequences are deduced is approximately 7000, and further, the number of full-length therein is only 5500. Thus, even including cDNA registered as EST, the percentage of human cDNA obtained so far is estimated to be 10-15% of all the genes.

[0006] It is possible to identify the transcription start site of mRNA on the genomic sequence based on the 5'-end sequence of a full-length cDNA, and to analyze factors involved in the stability of mRNA that is contained in the cDNA, or in its regulation of expression at the translation stage. Also, since a full-length cDNA contains ATG, the translation start site, in the 5'-region, it can be translated into a protein in a correct frame. Therefore, it is possible to produce a large amount of the protein encoded by the cDNA or to analyze biological activity of the expressed protein by utilizing an appropriate expression system. Thus, analysis of a full-length cDNA provides valuable information which complements the information from genome sequencing. Also, full-length cDNA clones that can be expressed are extremely valuable in empirical analysis of gene function and in industrial application.

[0007] Therefore, if a novel human full-length cDNA is isolated, it can be used for developing medicines for diseases in which the gene is involved. The protein encoded by the gene can be used as a drug by itself. Thus, it has great significance to obtain a full-length cDNA encoding a novel human protein.

[0008] In particular, human secretory proteins or membrane proteins would be useful by itself as a medicine like tissue plasminogen activator (TPA), or as a target of medicines like membrane receptors. In addition, genes for signal transduction-associated proteins (protein kinases, etc.), glycoprotein-associated proteins, transcription-associated proteins, etc. are genes whose relationships to human diseases have been elucidated. Moreover, genes for disease-associated proteins form a gene group rich in genes whose relationships to human diseases have been elucidated.

[0009] Therefore, it has great significance to isolate novel full-length cDNA clones of human, only few of which has been isolated. Especially, isolation of a novel cDNA clone encoding a secretory protein or membrane protein is desired since the protein itself would be useful as a medicine, and also the clones potentially include a gene associated with diseases. In addition, genes encoding proteins that are associated with signal transduction, glycoprotein, transcription, or diseases are expected to be useful as target molecules for therapy, or as medicines themselves. These genes form a gene group predicted to be strongly associated with diseases. Thus, identification of the full-length cDNA clones

encoding those proteins has great significance.

SUMMARY OF THE INVENTION

[0010] An objective of the present invention is to provide a polynucleotide encoding a novel protein, a protein encoded by said polynucleotide, and novel usages of these.

[0011] The inventors have developed a method for efficiently cloning a human full-length cDNA that is predicted by the ATGpr etc. to be a full-length cDNA clone, from a full-length-enriched cDNA library that is synthesized by the oligo-capping method. Then, the inventors determined the nucleotide sequence of the obtained cDNA clones from both 5'- and 3'- ends.

[0012] Furthermore, the inventors analyzed the obtained clones by the BLAST search of the databases, SwissProt (http://www.ebi.ac.uk/ebi_docsSwissProt_db/swisshome.html), GenBank (http://www.ncbi.nlm.nih.gov/web/GenBank), and UniGene (Human) (http://www.ncbi.nlm.nih.gov/UniGene).

[0013] The full-length cDNA clones of the present invention have high fullness ratio since these were obtained by the combination of (1) construction of a full-length-enriched cDNA library that is synthesized by the oligo-capping method, and (2) a system in which the full-length ratio is evaluated from the nucleotide sequence of the 5'-end (selection based on the ATGpr, previously removed complete sequences to ESTs). However, the primer of the present invention enables to obtain full-length cDNA easily without any specialized methods as in the described method.

Homology analysis in which the analysis is carried out against a not-full-length cDNA fragment to postulate the function of a protein encoded by said fragment, is being commonly performed.

However, since such analysis is based on the information of the fragment, it is not clear as to whether this fragment corresponds to a part that is functionally important in the protein. In other words, the reliability of the homology analysis based on the information of a fragment is doubtful, as information related to the structure of the whole protein is not available. However, the homology analysis of the present invention is conducted based on the information of a full-length cDNA comprising the whole coding region of the cDNA, and therefore, the homology of various portions of the protein can be analyzed. Hence, the reliability of the homology analysis has been dramatically improved in the present invention.

[0014] The inventors completed the invention by finding that it is possible to synthesize a novel full-length cDNA by using the combination of a primer that is designed based on the nucleotide sequence of the 5'-ends of the selected full-length cDNA clones and any of an oligo-dT primer or a 3'-primer that is designed based on the nucleotide sequence of the 3'-ends of the selected clones.

[0015] Thus, the present invention relates to primers described below, a method for synthesizing a polynucleotide using the primers, and polynucleotides obtained by the method.

[0016] First, the present invention relates to

(1) use of an oligonucleotide as a primer for synthesizing the polynucleotide comprising the nucleotide sequence set forth in any one of SEQ ID NOs: 1-5547 and SEQ ID NOs: 16111-16164, or the complementary strand thereof, wherein said oligonucleotide is complementary to said polynucleotide or the complementary strand thereof and comprises at least 15 nucleotides;

(2) a primer set for synthesizing polynucleotides, the primer set comprising an oligo-dT primer and an oligonucleotide complementary to the complementary strand of the polynucleotide comprising the nucleotide sequence set forth in any one of SEQ ID NOs: 1-5547 and SEQ ID NOs: 16111-16164, wherein said oligonucleotide comprises at least 15 nucleotides; and

(3) a primer set for synthesizing polynucleotides, the primer set comprising a combination of an oligonucleotide comprising a nucleotide sequence complementary to the complementary strand of the polynucleotide comprising a 5'-end nucleotide sequence and an oligonucleotide comprising a nucleotide sequence complementary to the polynucleotide comprising a 3'-end nucleotide sequence, wherein said oligonucleotides comprise at least 15 nucleotides and wherein said combination of 5'-end nucleotide sequence / 3'-end nucleotide sequence is selected from the combinations of 5'-end nucleotide sequence / 3'-end nucleotide sequence set forth in the SEQ ID NOs in Tables 1 and 2.

[0017] Tables 1 and 2 shows names of clones obtained in the examples described later, comprising the polynucleotide of the present invention (Table 1; 5547 clones, Table 2; 54 clones), names of nucleotide sequences at the 5'-end and 3'-end of the full-length cDNA, and their corresponding SEQ ID NOs. A blank indicates that the 3'-end sequence corresponding to the 5'-end sequence has not been determined for the same clone.

[0018] The SEQ ID NO of a 5'-end sequence is shown on the right side of the name of the 5'-end sequence, and the SEQ ID NO of a 3'-end sequence is shown on the right side of the name of the 3'-end sequence.